Cambrian AI Research
explore the explosion

2022 AI HW Summit: The Annual Top 10 List

Karl Freund
Founder and Principal Analyst
Cambrian-AI Research LLC

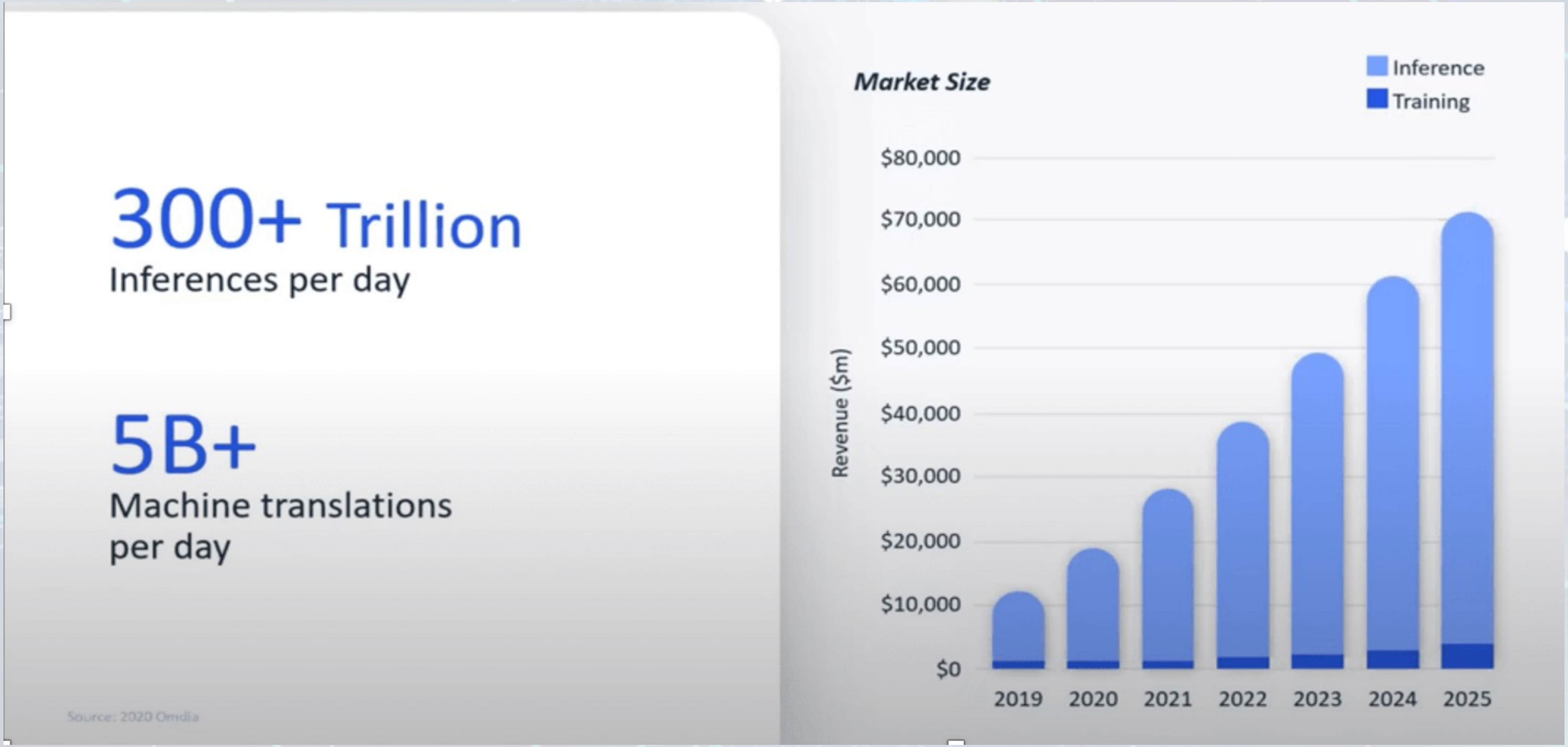
Karl.Freund@Cambrian-ai.com
@karlfreund

The 2021 Top 10 List

➔ All are still Evolving

1. MLPerf Benchmarks
2. Synopsys DSO.ai
3. Sambanova, Groq, & Tenstorrent
4. AWS: Inferentia
5. Google: TPU-V4
6. Intel Habana Labs Gaudi
7. Graphcore 2nd Generation and the IPU-Machine
8. NVIDIA Grace
9. Cerebras WSE-2 and Brain-Scale AI
10. …. (The founding or Cambrian-AI, of course!)

Why rob Banks? That’s where the money is.



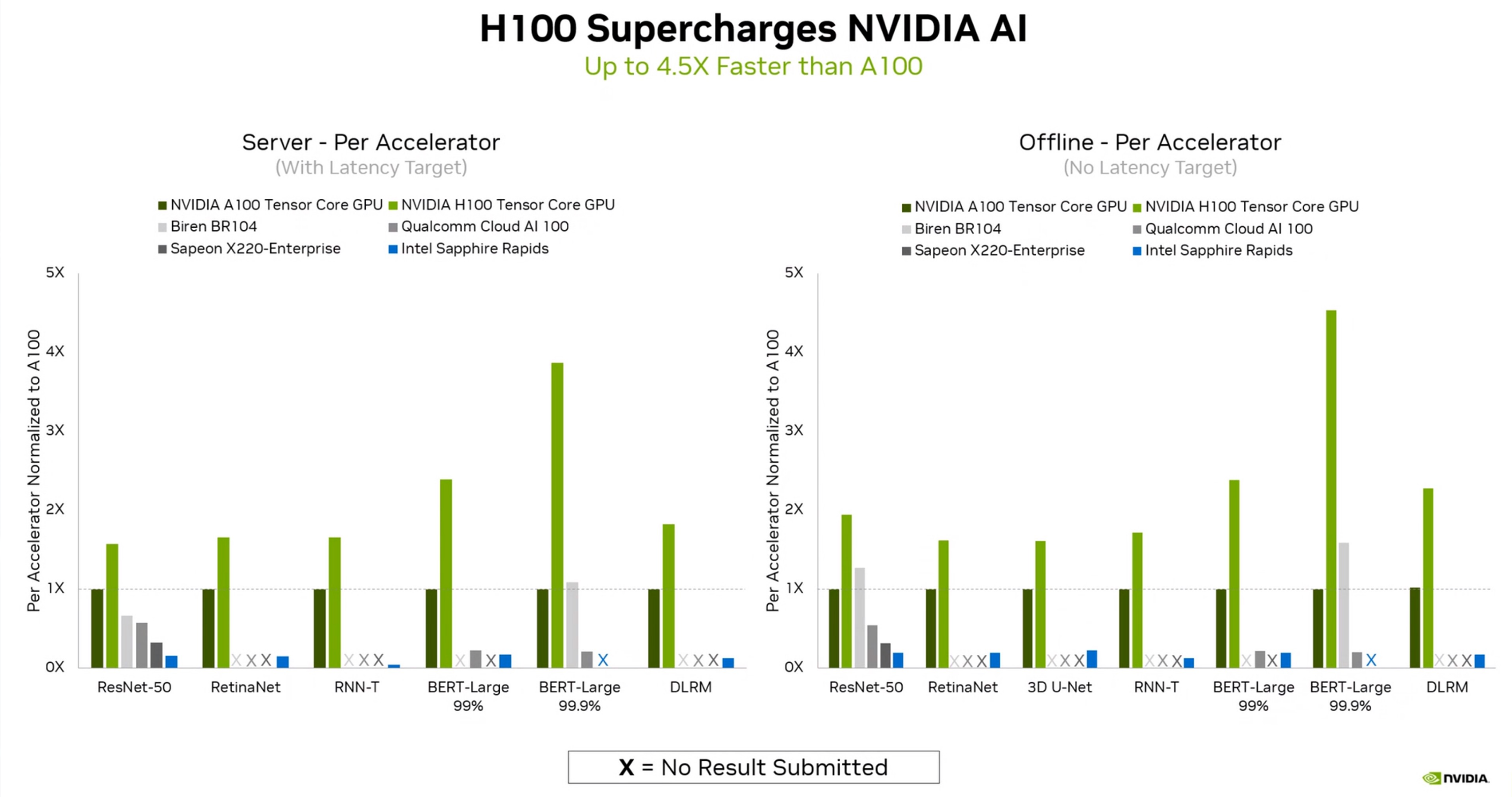
300+ Trillion
Inferences per day

5B+
Machine translations per day

Source: 2020 Omdia

MLPerf Inference 2.1

Cambrian AI Research
explore the explosion

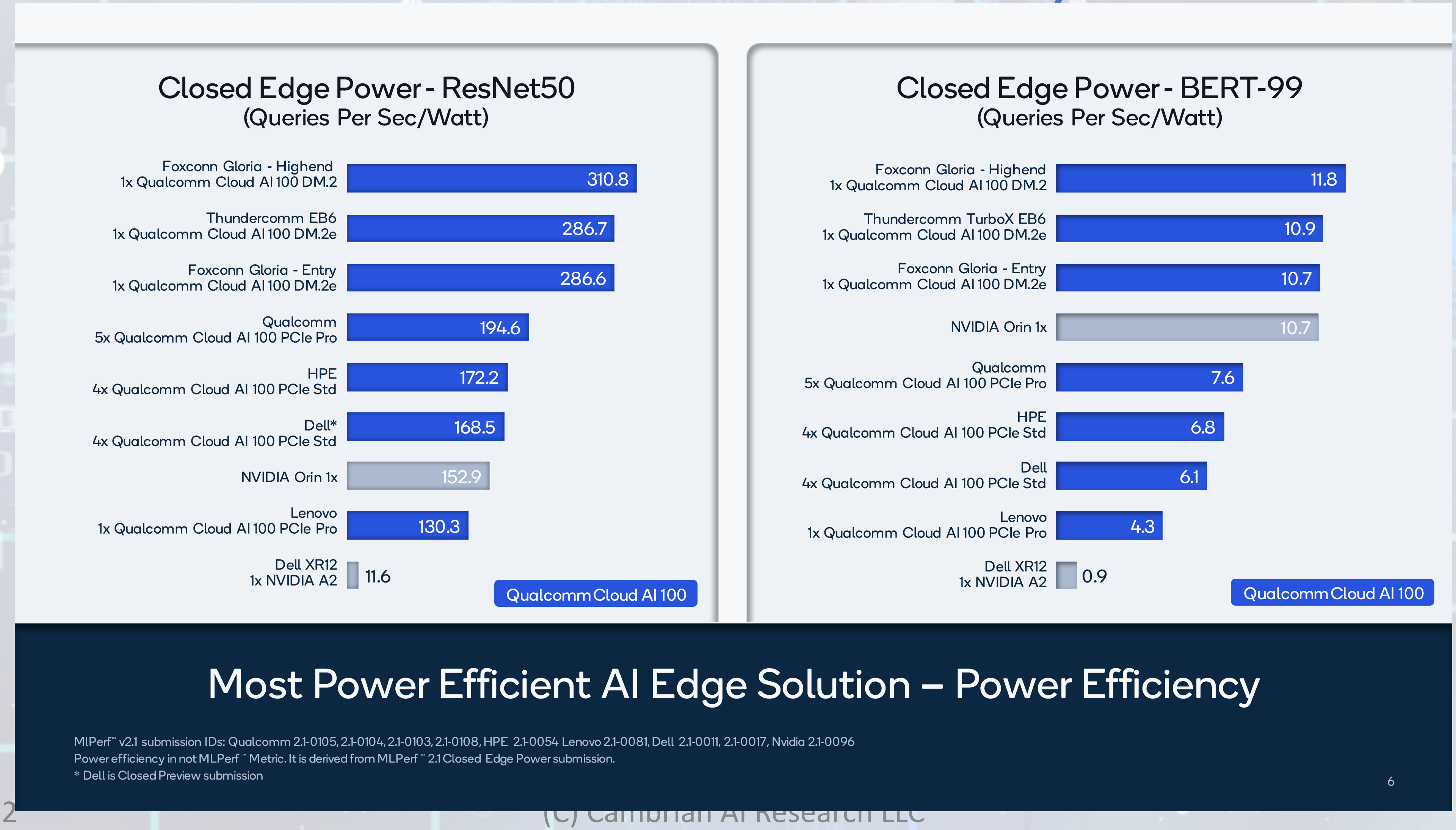


H100 Supercharges NVIDIA AI

Up to 4.5X Faster than A100

Qualcomm Leads MLPerf Efficiency

Over 200 submissions with Server Ecosystem



Closed Edge Power - ResNet50
(Queries Per Sec/Watt)

System	Value
Foxconn Gloria - Highend 1x Qualcomm Cloud AI 100 DM.2	310.8
Thundercomm EB6 1x Qualcomm Cloud AI 100 DM.2e	286.7
Foxconn Gloria - Entry 1x Qualcomm Cloud AI 100 DM.2e	286.6
Qualcomm 5x Qualcomm Cloud AI 100 PCIe Pro	194.6
HPE 4x Qualcomm Cloud AI 100 PCIe Std	172.2
Dell* 4x Qualcomm Cloud AI 100 PCIe Std	168.5
NVIDIA Orin 1x	152.9
Lenovo 1x Qualcomm Cloud AI 100 PCIe Pro	130.3
Dell XR12 1x NVIDIA A2	11.6

Qualcomm Cloud AI 100

Closed Edge Power - BERT-99
(Queries Per Sec/Watt)

System	Value
Foxconn Gloria - Highend 1x Qualcomm Cloud AI 100 DM.2	11.8
Thundercomm TurboX EB6 1x Qualcomm Cloud AI 100 DM.2e	10.9
Foxconn Gloria - Entry 1x Qualcomm Cloud AI 100 DM.2e	10.7
NVIDIA Orin 1x	10.7
Qualcomm 5x Qualcomm Cloud AI 100 PCIe Pro	7.6
HPE 4x Qualcomm Cloud AI 100 PCIe Std	6.8
Dell 4x Qualcomm Cloud AI 100 PCIe Std	6.1
Lenovo 1x Qualcomm Cloud AI 100 PCIe Pro	4.3
Dell XR12 1x NVIDIA A2	0.9

Qualcomm Cloud AI 100

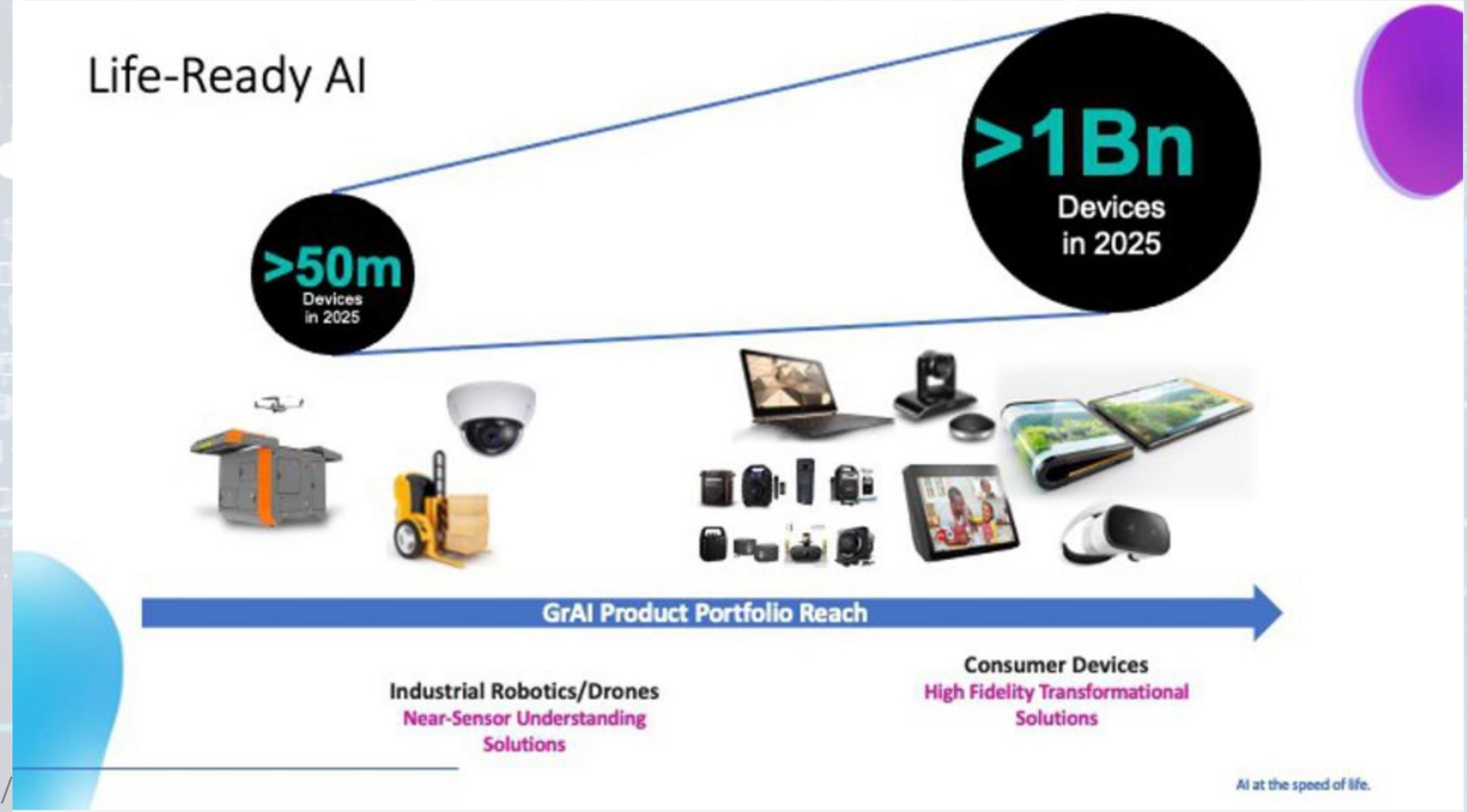
Most Power Efficient AI Edge Solution – Power Efficiency

MlPerf™ v2.1 submission IDs: Qualcomm 2.1-0105, 2.1-0104, 2.1-0103, 2.1-0108, HPE 2.1-0054 Lenovo 2.1-0081, Dell 2.1-0011, 2.1-0017, Nvidia 2.1-0096
Power efficiency in not MLPerf™ Metric. It is derived from MLPerf™ 2.1 Closed Edge Power submission.
* Dell is Closed Preview submission

The Top 10 AI HW Innovations of 2022

1. **GrAI Matter Labs** – High Fidelity Edge processing
2. **D-Matrix:** In-Memory Computation
3. **Untether.AI:** At-Memory Computation
4. **Mythic:** Here comes Analog compute
5. **Esperanto** : 1000 RISC-V cores
6. **SimaAI:** MLSoC provides SW-Centric ML
7. **AMD**: Massive FLOPS w/ A100-class TOPS
8. **Graphcore**: Wafer on Wafer, and Good Computer
9. **Intel Habana Labs**: Gaudi 2 doubles NVIDIA A100 throughput
10. **NVIDIA** Hopper and Grace **Superchips**: the Next Generation of Compute

GrAI Matter Labs:
High Fidelity real-time inference



Life-Ready AI

>50m Devices in 2025

>1Bn Devices in 2025

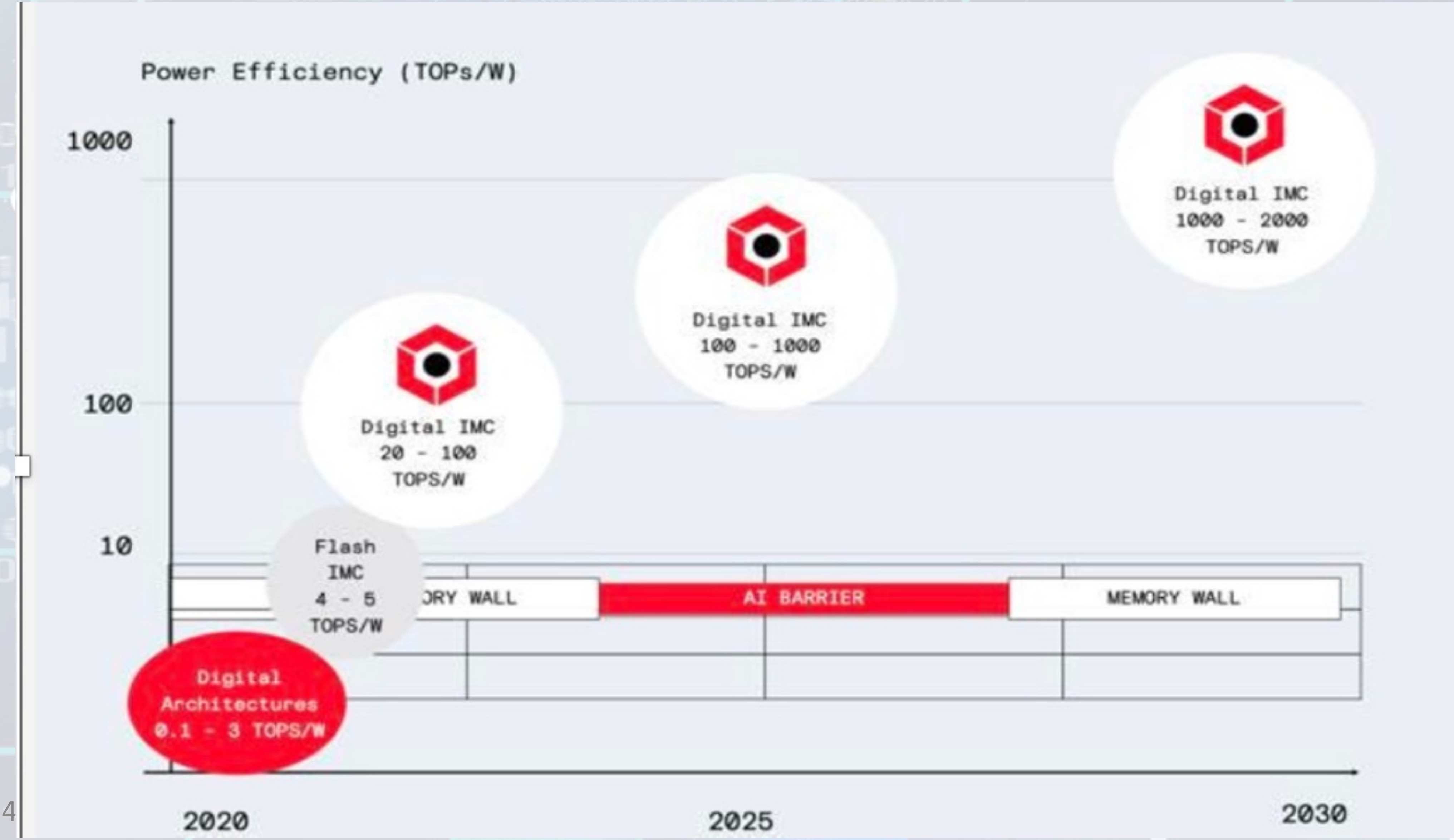
GrAI Product Portfolio Reach

Industrial Robotics/Drones
Near-Sensor Understanding Solutions

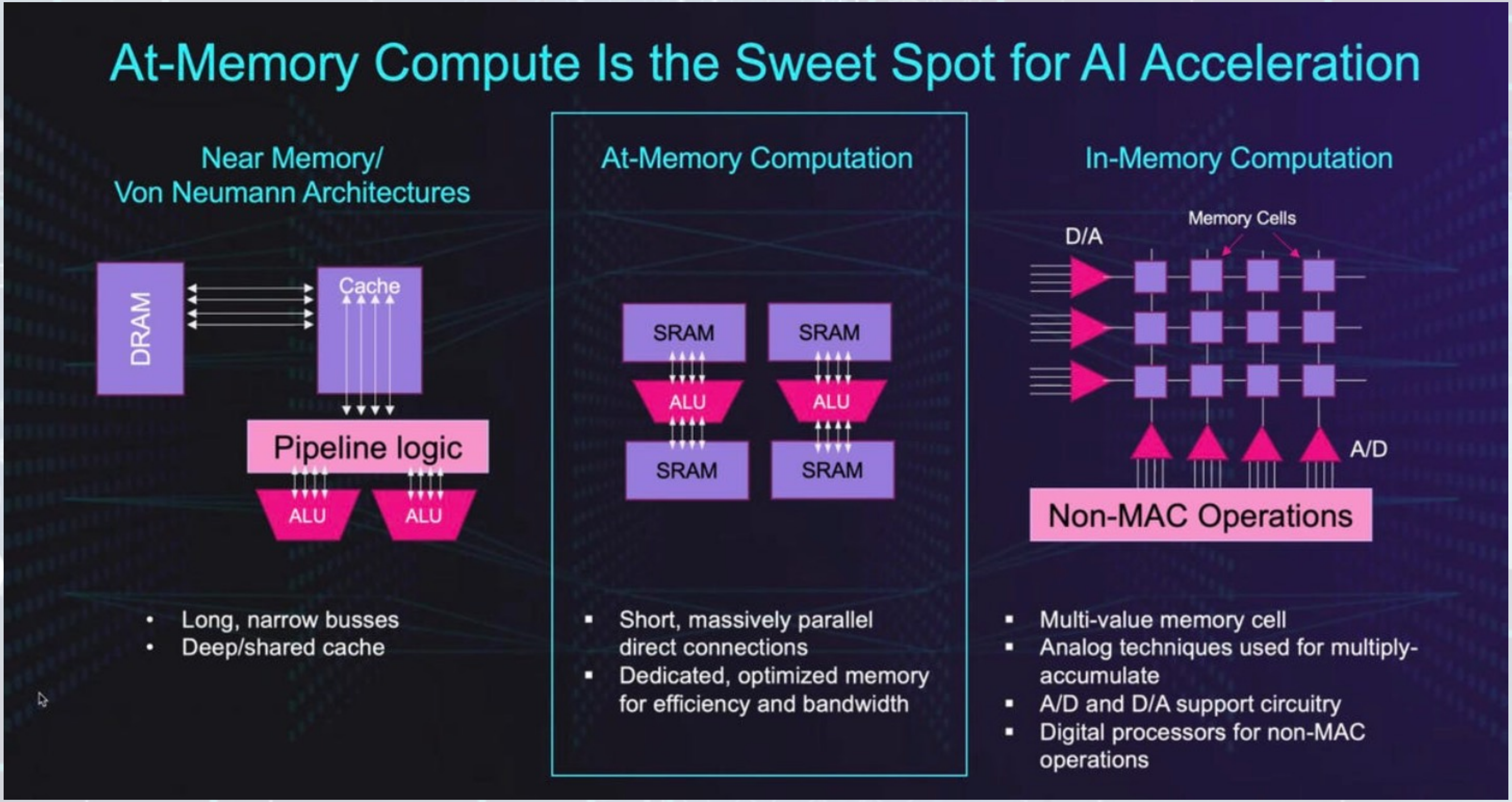
Consumer Devices
High Fidelity Transformational Solutions

AI at the speed of life.

D-Matrix: In-Memory Compute



Untether AI: 30 TFLOPS/Watt



At-Memory Compute Is the Sweet Spot for AI Acceleration

Near Memory/ Von Neumann Architectures

- Long, narrow busses
- Deep/shared cache

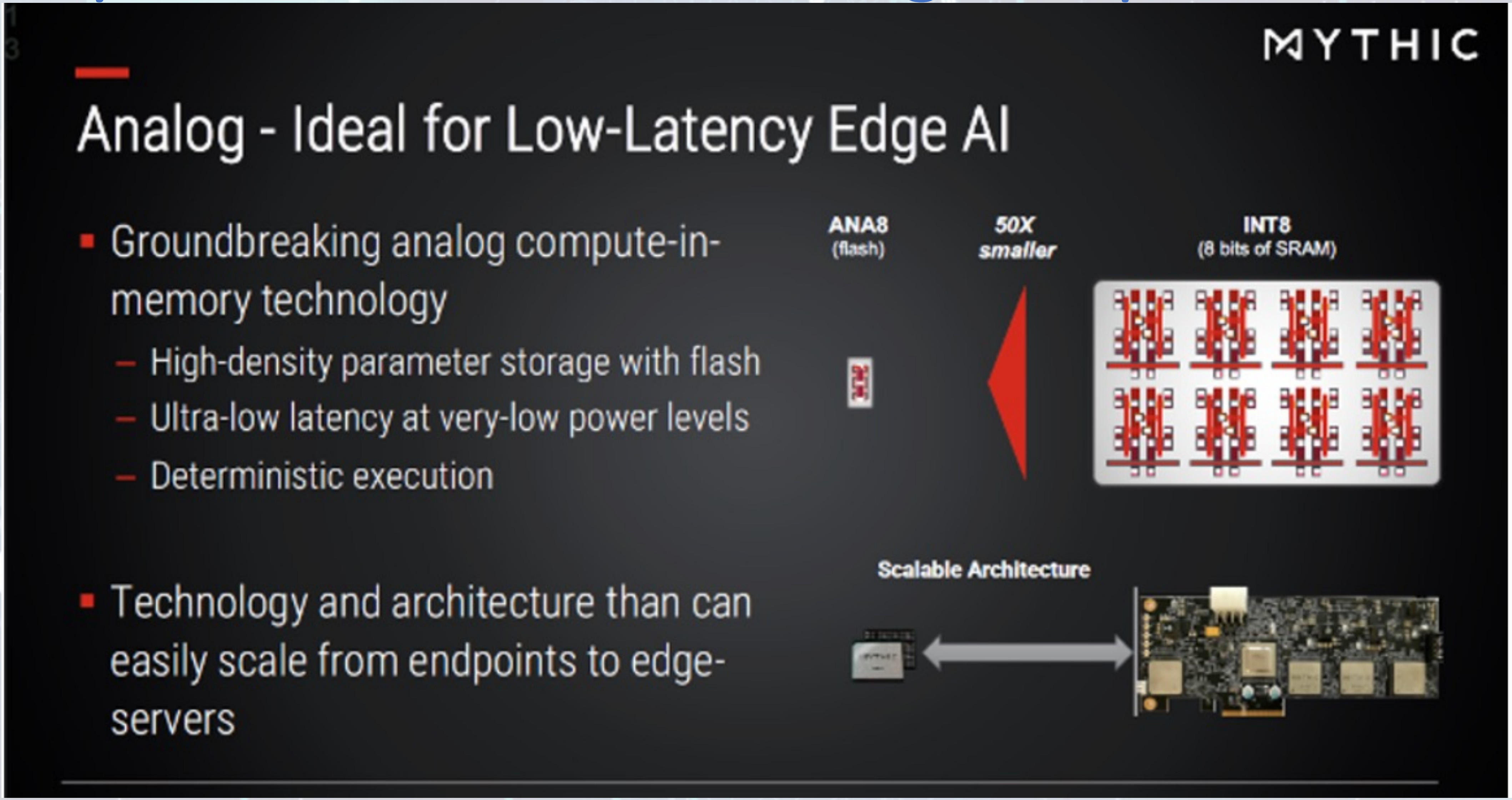
At-Memory Computation

- Short, massively parallel direct connections
- Dedicated, optimized memory for efficiency and bandwidth

In-Memory Computation

- Multi-value memory cell
- Analog techniques used for multiply-accumulate
- A/D and D/A support circuitry
- Digital processors for non-MAC operations

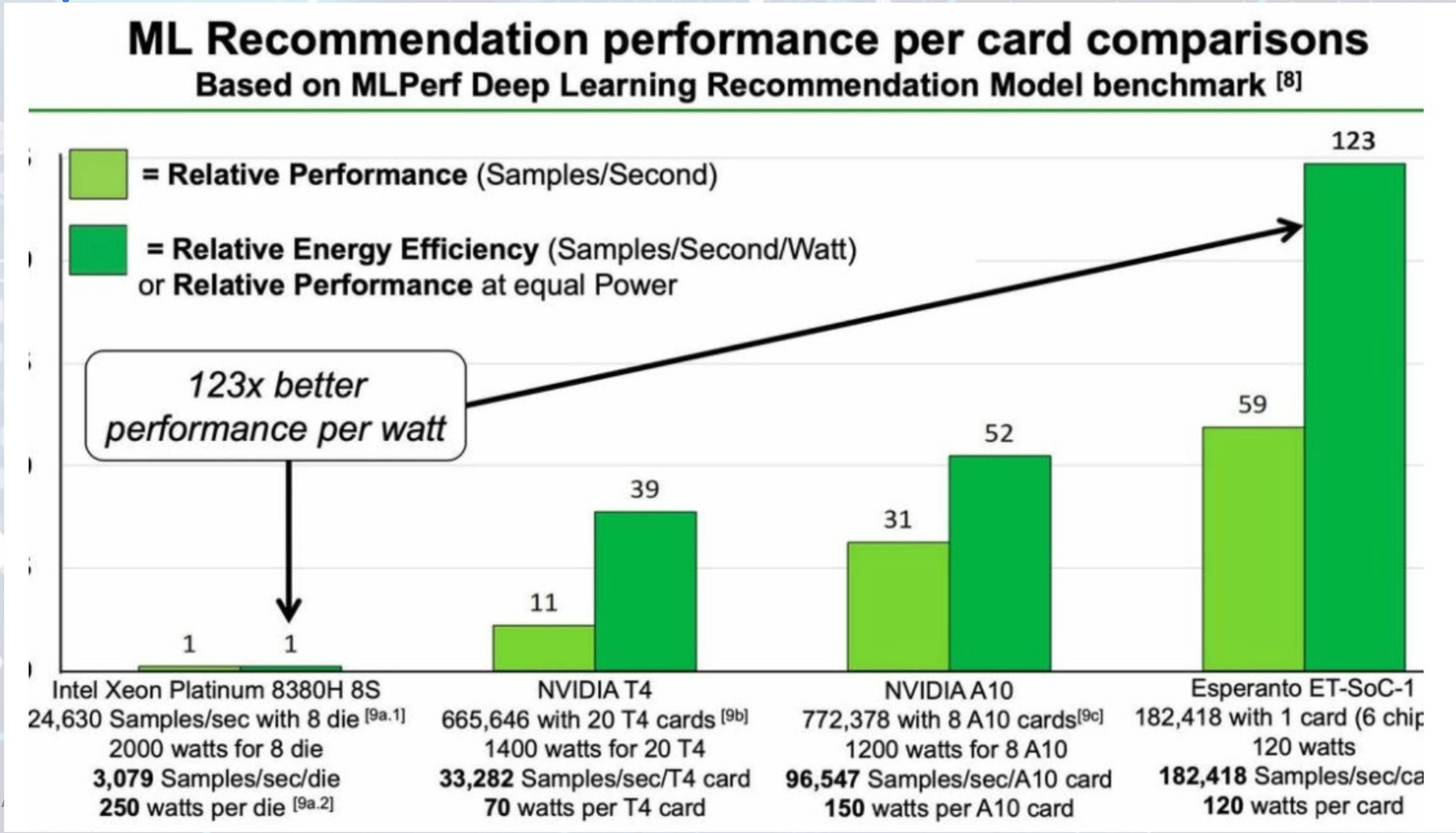
Mythic: Here comes Analog Compute



Analog - Ideal for Low-Latency Edge AI

- Groundbreaking analog compute-in-memory technology
  - High-density parameter storage with flash
  - Ultra-low latency at very-low power levels
  - Deterministic execution
- Technology and architecture than can easily scale from endpoints to edge-servers

Esperanto: 1000 RISC-V Cores



ML Recommendation performance per card comparisons

Based on MLPerf Deep Learning Recommendation Model benchmark [8]

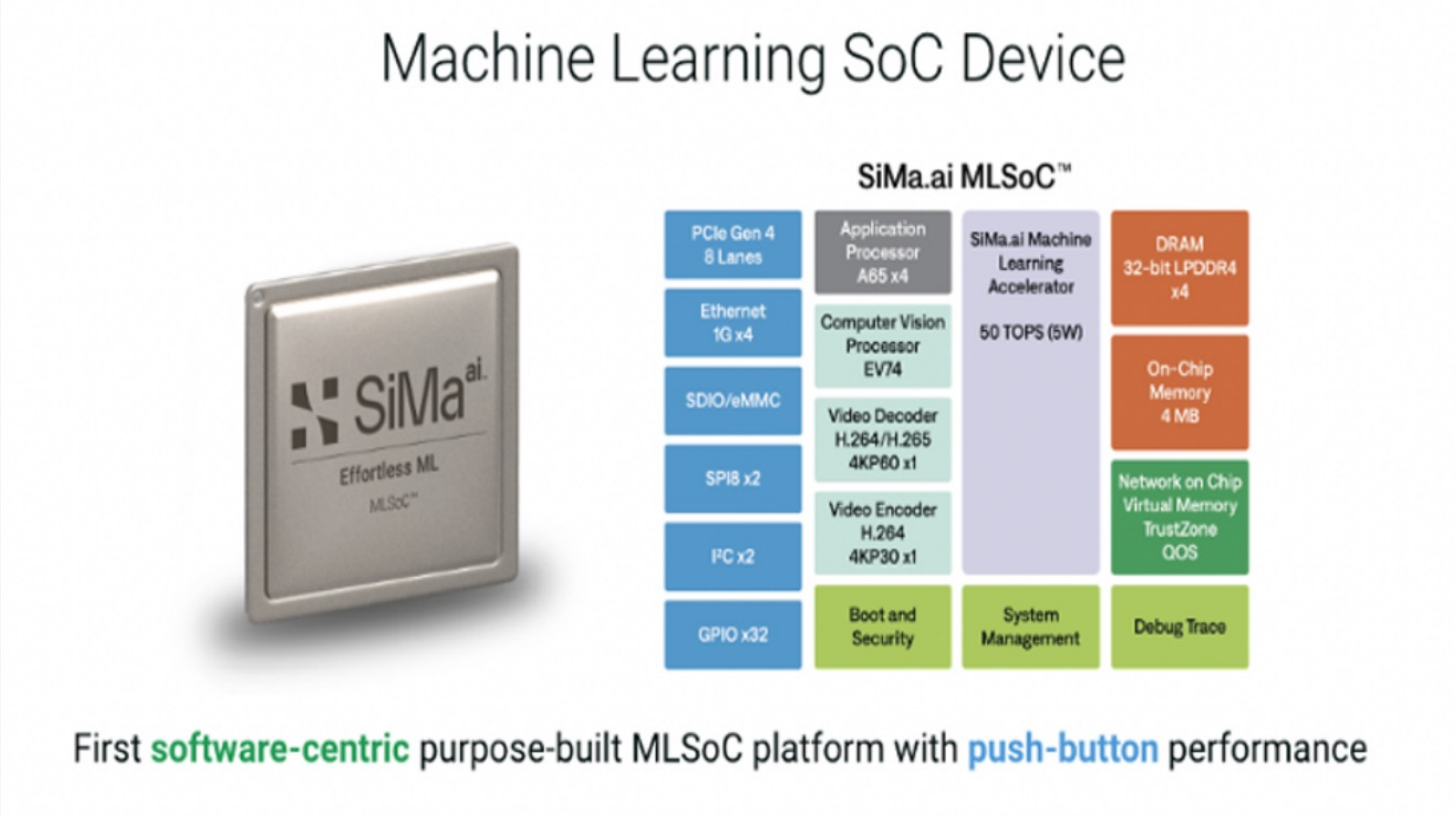
Intel Xeon Platinum 8380H 8S
24,630 Samples/sec with 8 die [9a.1]
2000 watts for 8 die
3,079 Samples/sec/die
250 watts per die [9a.2]

NVIDIA T4
665,646 with 20 T4 cards [9b]
1400 watts for 20 T4
33,282 Samples/sec/T4 card
70 watts per T4 card

NVIDIA A10
772,378 with 8 A10 cards[9c]
1200 watts for 8 A10
96,547 Samples/sec/A10 card
150 watts per A10 card

Esperanto ET-SoC-1
182,418 with 1 card (6 chip
120 watts
182,418 Samples/sec/ca
120 watts per card

SiMa.ai MLSoC



Machine Learning SoC Device

SiMa.ai MLSoC™

PCIe Gen 4 8 Lanes

Ethernet 1G x4

SDIO/eMMC

SPI8 x2

I²C x2

GPIO x32

Application Processor A65 x4

Computer Vision Processor EV74

Video Decoder H.264/H.265 4KP60 x1

Video Encoder H.264 4KP30 x1

Boot and Security

SiMa.ai Machine Learning Accelerator

50 TOPS (5W)

System Management

DRAM 32-bit LPDDR4 x4

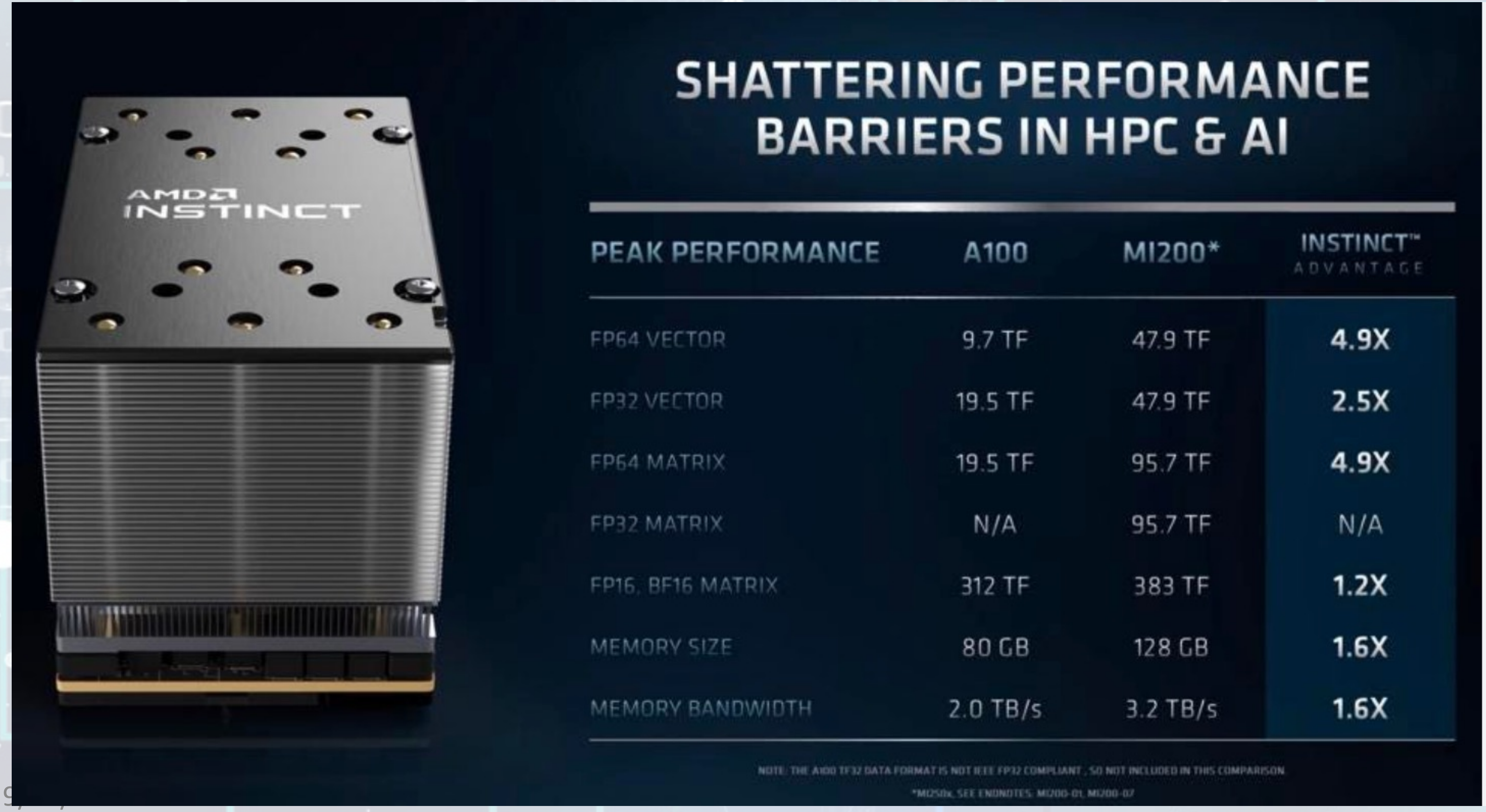
On-Chip Memory 4 MB

Network on Chip Virtual Memory TrustZone QOS

Debug Trace

First software-centric purpose-built MLSoC platform with push-button performance

AMD Instinct MI200



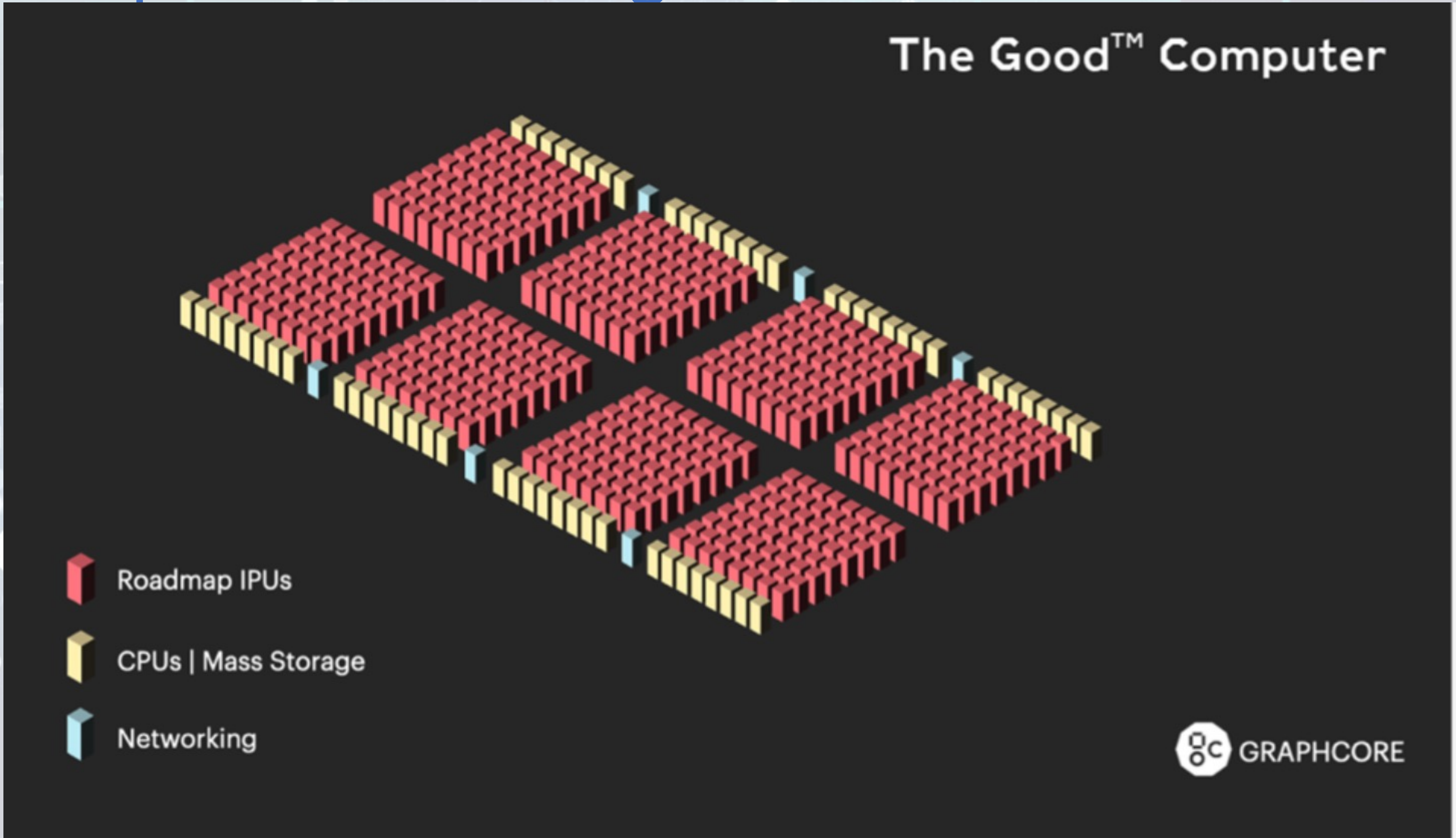
SHATTERING PERFORMANCE BARRIERS IN HPC & AI

PEAK PERFORMANCE	A100	MI200*	INSTINCT™ ADVANTAGE
FP64 VECTOR	9.7 TF	47.9 TF	4.9X
FP32 VECTOR	19.5 TF	47.9 TF	2.5X
FP64 MATRIX	19.5 TF	95.7 TF	4.9X
FP32 MATRIX	N/A	95.7 TF	N/A
FP16, BF16 MATRIX	312 TF	383 TF	1.2X
MEMORY SIZE	80 GB	128 GB	1.6X
MEMORY BANDWIDTH	2.0 TB/s	3.2 TB/s	1.6X

NOTE: THE A100 TF32 DATA FORMAT IS NOT IEEE FP32 COMPLIANT , SO NOT INCLUDED IN THIS COMPARISON

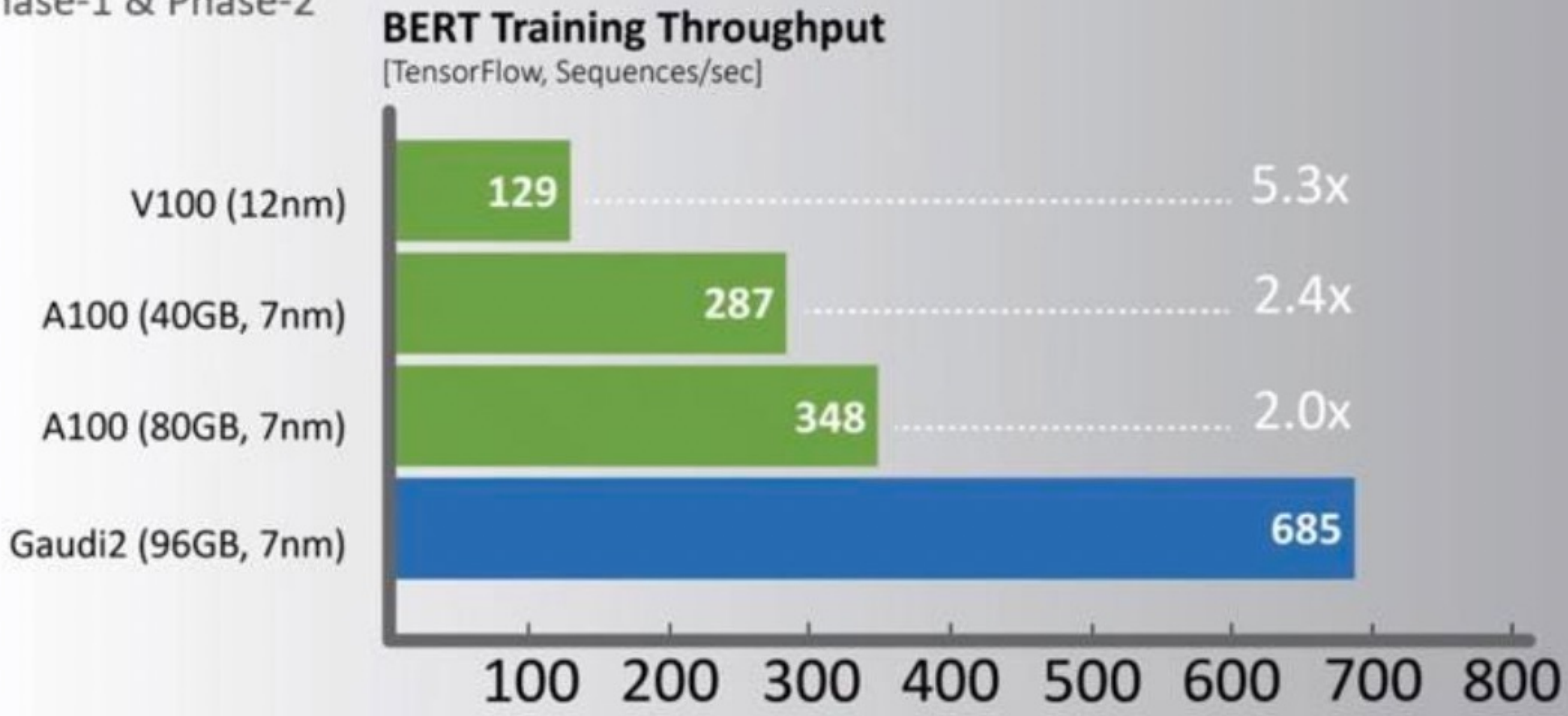
*MI250x, SEE ENDNOTES: MI200-01, MI200-07

Graphcore: Thinking BIG

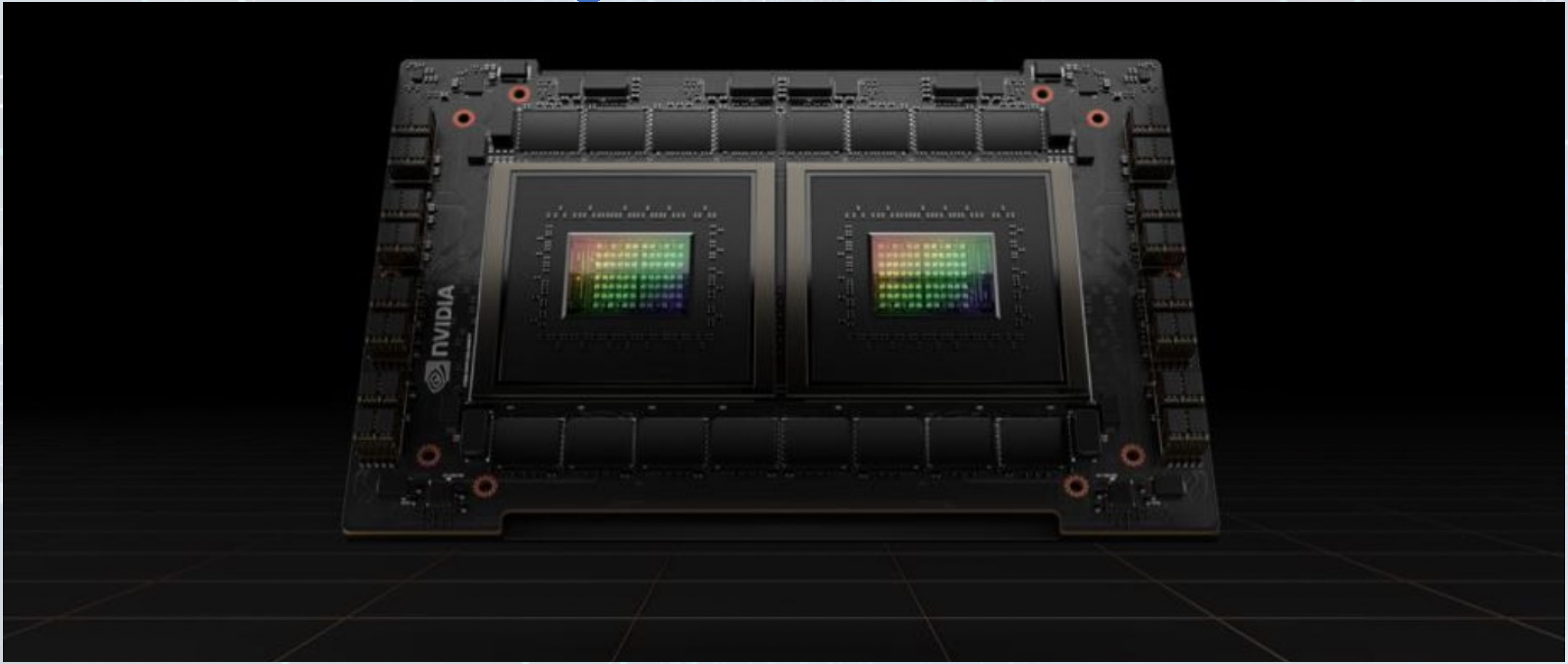


Intel Habana Labs Gaudi2: 2X A100

Effective Throughput:
Combining Phase-1 & Phase-2



Grace Superchips:
The Future of High Performance



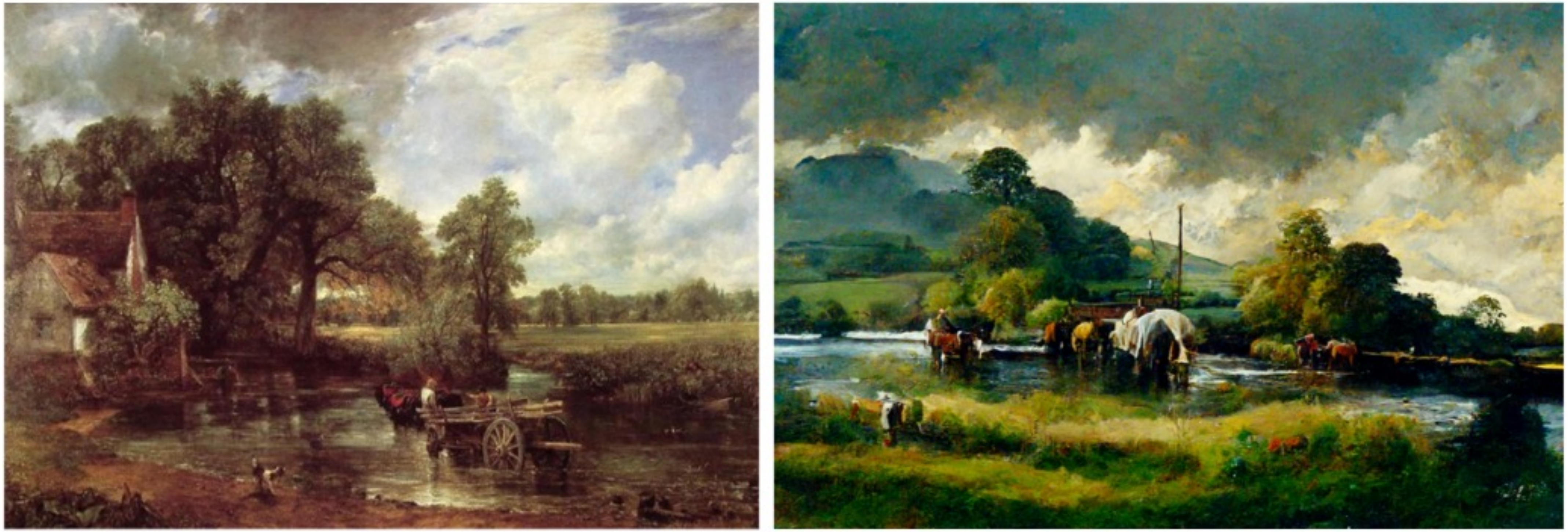
9/14/22

Closing Thoughts

- NVIDIA is practically unassailable in data center training, where Intel, Cerebras, Graphcore, SambaNova, Tenstorrent, and Groq are all attacking.
  - So, if you want to go after this space, you need magic tech and a lot of money
- The Edge market(s) are ripe with opportunities with many niches.
- Software matters even more than you think
  - NVIDIA improved Jetson by **50%** last MLPerf round
- BIG networks (LLMs) will eat the world
  - NVIDIA
  - Graphcore
  - Cerebras

LLMs being used to create images from text

Prompt: “oil on canvas painting + romanticism + landscape + a hay wain pulled by two horses as it crosses a river + a backdrop of mountains, trees, and clouds in the background + simple and idyllic depiction of rural life in England”



Left: The Hay Wain by John Constable (Public Domain). Right: Image by Alberto Romero via Midjourney

THANK YOU!

Have a great conference!

Our clients include:

cerebras	Esperanto TECHNOLOGIES	GRAPHCORE	IBM	intel.	NVIDIA.
Qualcomm	Simple Machines	SYNOPSYS	Tenstorrent	SiFive	TEMASEK
d-Matrix	GML GrAI Matter Labs	KeyBank	MYTHIC	SiMa ai.	VENTANA MICRO

Come visit our site for news and research reports
http://www.Cambrian-ai.com